Protein Engineering vol.14 no.7 pp.459–463, 2001

A numerical measure of amino acid residues similarity based on the analysis of their surroundings in natural protein sequences

Sergey I.Rogov and Alexei N.Nekrasov[1]

Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Miklukho-Maklaya St. 16/10, GSP-7, Moscow 117997, Russia

[1]To whom correspondence should be addressed.
E-mail: alexei_nekrasov@mail.ru

A measure of similarity between amino acid residues based on the analysis of the surroundings of each residue in primary structures of native proteins is proposed. The statistical data used for this purpose were obtained from the analysis of 168,808 protein sequences, which comprise the Protein Identification Research database (release 63). Using various threshold values of the proposed measure, amino acid residues were classified into several groups. The classification elaborated differs essentially from groupings previously used. The numerical measure of amino acid residues similarity can be used in site-directed mutagenesis studies for the prediction of probability of local spatial rearrangements in proteins.

Keywords: amino acid classification/database/protein sequences

Introduction

Nowadays, one of the main approaches in protein research is modification of proteins by genetic methods. The key element in these studies is the production of new recombinant proteins and their modification aimed at alteration of their biological activity. The method of point mutations is most frequently used for this purpose (Hurley *et al.*, 1992; Lim *et al.*, 1992; Zhang *et al.*, 1992). It is often of crucial importance to preserve the structure of the modified protein akin to its native conformation while altering its substrate specificity or affinity to the regulatory factors. As a rule, the experimental confirmation of the equivalence between three-dimensional structures of the native and recombinant proteins is time-consuming. Hence, the necessity arises to predict the influence of amino acid substitutions upon protein structure. Various types of classifications of amino acid residues are now used to solve this problem (Johnson and Overington, 1986; Taylor, 1986; Bordo and Argos, 1991; Topham *et al.*, 1997; Murphy *et al.*, 2000). The approaches used fall into two major types: the first is based on measurement or evaluation of various physical-chemical properties of amino acid residues (Taylor, 1986); the second on the analysis of amino acid substitutions in families of evolutionary related proteins (Bordo and Argos, 1991; Topham *et al.*, 1997; Murphy *et al.*, 2000). In our opinion, both approaches suffer from inherent drawbacks. A certain degree of arbitrariness in selection of physical-chemical properties of the residues and methods of their determination is inherent to the first of the above-mentioned approaches. Thus, various authors (Janin, 1979; Wolfenden *et al.*, 1981; Kyte and Doolite, 1982; Rose *et al.*, 1985) report the residue hydrophobicity classifications, which differ considerably from each other. The main disadvantage of methods based on comparison of the frequency of the amino acid substitutions is that the probability of substitution of a given residue depends on its role in protein structure or function. Since various families of proteins have different folds, the probability of substitution of a given residue for any other will vary for different families. Thus, classifications of the residues, based on such an approach, will depend on what family of proteins was analysed.

We believe it would be more correct to introduce a continuous numerical criterion based on the analysis of residue surroundings in protein primary structures. To disclose the correlation between the physical-chemical properties of the residues and their surrounding in the sequence is not only important for protein engineering, but could also be used for deducing the protein structure from the sequence. We consider that in most cases, similarity of residues surroundings reflects similar structural features of their local architecture. This must be apparent for a large set of sequences, where individual traits of the protein families are even. Accordingly, substitution of a given residue by another with similar surroundings is likely to result in preservation of the local spatial architecture. Thus, the numerical measure of the similarity surroundings of amino acid residues in primary structures of proteins would allow a classification of residues which differs from that which is currently used and, simultaneously a numerical criterion of influence of the amino acid substitutions upon the local structure. In this paper an attempt to introduce such a criterion is undertaken.

Materials and methods

Data

The statistical data were obtained using 168,808 native protein sequences included in the Protein Identification Research (PIR) database (release 63). The total number of amino acid residues in the sequences considered was 58,112,946. All available protein sequences without any preliminary selection were used as a primary data set. Such an approach allows one to eliminate specific features of individual primary structures and to reveal regularities, intrinsic to all native amino acid sequences.

Analysis of the PIR database

The study included the following stages: (i) reconstruction of averaged surroundings in protein sequences for each of 20 amino acid residues; (ii) determination of the characteristic length of a sequence segment with the most pronounced mutual influence of amino acid residues; and (iii) comparison of the amino acid residue surroundings in primary structures of native proteins.

First, the total number of all 400 pairs of amino acid residues separated by i peptide bonds [$N(i)$] was calculated. For each X–Z pair, the values of its absolute $N_{XZ}(i)$ and relative $c_{XZ}(i)$ content in the database were determined:

$$c_{XZ}(i) = \frac{N_{XZ}(i)}{N(i)} \quad (1)$$

Let us consider the distribution of relative content of residue Z in the neighbourhood of residue X separated by 1 to n peptide bonds. Let i be positive if Z is closer to the C-terminus of polypeptide chain than X, and negative if otherwise. At the first stage, the value n was set to 55. In the given neighbourhood the average relative content of Z equals

$$C_{XZ} = \frac{\sum_{i=1}^{n} (c_{XZ}(i) + c_{XZ}(-i))}{2n} \quad (2)$$

Let us consider the function $d_{XZ}(i)$, which represents the normalised deviation of the relative content of X–Z pairs separated by i peptide bonds, from the average:

$$d_{XZ}(i) = \frac{c_{XZ}(i) - C_{XZ}}{C_{XZ}} \quad (3)$$

This function can be interpreted as a distribution of relative content of Z (hereafter referred to as the distributed residue) in the neighbourhood of X (hereafter referred to as the central residue).

To evaluate the characteristic size of a sequence fragment, within which the pronounced difference of the content of pairs of amino acid residues from average values is observed, the value of the root mean square deviation $s(i)$ from 0 in a sample of 400 d_{XZ} values for all pairs of residues was used. Its distribution against i is as follows:

$$s(i) = \frac{\sqrt{\sum_{j=1}^{20} \sum_{k=1}^{20} (d_{X_j Z_k}(i))^2}}{400} \quad (4)$$

The numerical measure of residues surroundings similarity was determined as follows. Let $d_{X_1 Z_k}$ and $d_{X_2 Z_k}$ be the known distributions of residues Z_k in the neighbourhood of the residues X_1 and X_2, $k = 1, 2, \ldots 20$. Then, the sum of distances between vectors $d_{X_1 Z_k}$ and $d_{X_2 Z_k}$ is calculated as follows:

$$r_{X_1 X_2} = \sum_{k=1}^{20} \sqrt{\sum_{i=1}^{n} \left((d_{X_1 Z_k}(i) - d_{X_2 Z_k}(i) \right)^2 + \left(d_{X_1 Z_k}(-i) - d_{X_2 Z_k}(-i) \right)^2} \quad (5)$$

The mean distance between vectors d_{XZ} for all possible pairs of residues is calculated as follows:

$$R = \frac{\sum_{j=1}^{20} \sum_{k=j}^{20} r_{X_j X_k}}{\sum_{i=1}^{20} i} = \frac{\sum_{j=1}^{20} \sum_{k=j}^{20} r_{X_j X_k}}{210} \quad (6)$$

The following value has been introduced as a measure of similarity of environments of residues X_1 and X_2:

$$m_{X_1 X_2} = 1 - \frac{r_{X_1 X_2}}{R} \quad (7)$$

Equipment and software

The calculations were made using the original software written in C++ for an IBM PC-compatible computer.

Results and discussion

The characteristic diagrams of distribution of d_{XZ} against i are plotted in Figure 1. The common feature of these distributions irrespective of any given X–Z pair is the decrease in variance of the relative content of the distributed residue with the increase of the number of peptide bonds between the residues.

It should be noted that when the distributed and central residues are identical, the diagrams of distribution are symmet-

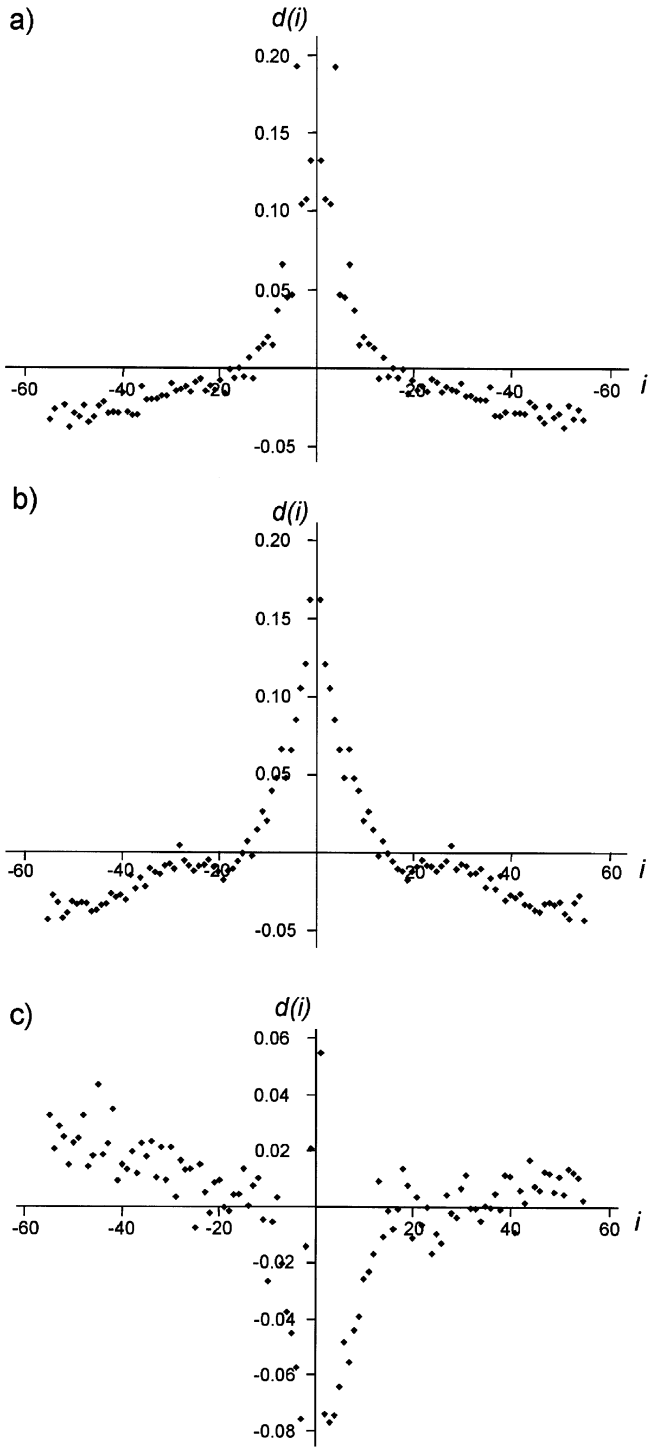

Fig. 1. Distribution of the relative content d of one amino acid residue in the neigbourhood of another against the number of peptide bonds between them (i). (**a**) –**A** in the neighbourhood of **A**; (**b**) –**R** in the neighbourhood of **R**; (**c**) –**P** in the neighbourhood of **D**.

ric. The form of the branches on the diagrams is close to exponential (Figures 1a and b). These results are in good agreement with the previously reported data (Poroykov *et al*., 1976) on increased probability of a common grouping of identical residues in a polypeptide chain. When the central and distributed residues are different, the form of the diagram may differ significantly from exponential (Figure 1c).

The diagram of $s(i)$ is presented in Figure 2. The result demonstrates that the most pronounced mutual influence of the residues is observed when the number of peptide bonds between them does not exceed 20. It should be noted that for a number of pairs (A–A, R–R and others) the mutual influence remains significant even on distances exceeding 50 peptide bonds between the residues. According to the data previously reported (Cserzo and Simon, 1989), the maximal distance of mutual influence was determined to be about nine peptide bonds. It is noteworthy in this context that a local minimum in distribution of the root mean square deviation was observed at $i = 9$ (Figure 2). Probably, it reflects a certain level of protein spatial organisation. The presence of this minimum has probably led the authors (Cserzo and Simon, 1989) to make the conclusion about the primary role of interactions within short segments of polypeptide chain in formation of spatial structure of proteins.

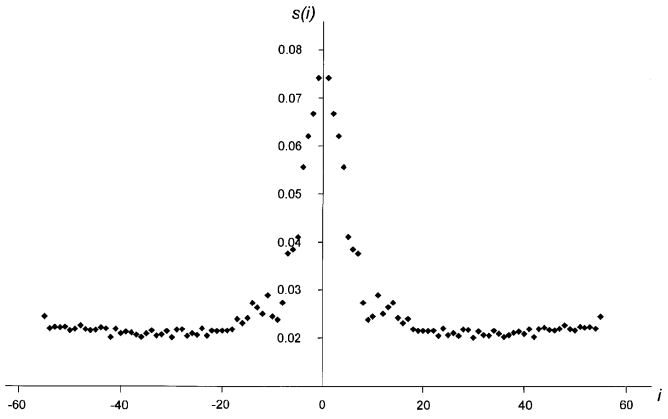

Fig. 2. Dependence of $s(i)$ (Equation 4) on the number of peptide bonds i between residues.

The data suggest that the mutual influence of amino acid residues is not limited to the nearest neighbours, but extends across significant distances in a polypeptide chain. Therefore we used an interval from 1 to 20 peptide bonds for the comparison of surroundings of amino acid residues.

Values of m for all 400 possible pairs of residues were calculated according to Equations (2), (3), (5), (6) and (7) with $n = 20$. The corresponding numbers are shown in Table I. According to Equation (6), m are nearer to 1 for those residues whose surroundings display a higher degree of similarity. The increase in dissimilarity of the residues surroundings corresponds to a decrease in m.

As was noted in the Introduction, the proposed measure of similarity of surroundings of the amino acid residues in primary structures of native proteins is a continuous numerical criterion. However, using various threshold values of m, it is possible to allocate groups of residues with an appropriate level of similarity of surroundings. The conformity of these groups of residues to earlier classifications is of particular interest. We have introduced the following threshold values of m: 0.4, 0.3, 0.2 and 0.1.

With the threshold value $m = 0.4$ only the string **V–I–L** and the **S–T** pair can be detected among all residues. According to the earlier classification (Taylor, 1986), the first three residues comprise a group of non-polar residues with aliphatic side chains. Also, these residues are grouped together in the classifications based on the analysis of amino acid substitutions (Bordo and Argos, 1991; Murphy *et al*., 2000). The main feature of these residues is the high degree of hydrophobicity.

High degree of similarity of surroundings of residues **S** and **T** can be accounted for by likeliness of structure and properties of their side chains: the small size and the ability to form hydrogen bonds are common for both residues. The similarity between the side chains of **S** and **T** has been noted in all classifications (Taylor, 1986; Bordo and Argos, 1991; Johnson and Overington, 1993; Topham *et al*., 1997; Murphy *et al*., 2000). However these residues never formed the center of the separate group.

With the reduction of m threshold value to 0.3, the group of the hydrophobic residues incorporates **F** and **Y**. It should be noted that **F** appears to be closer to **I** ($m = 0.380$), rather

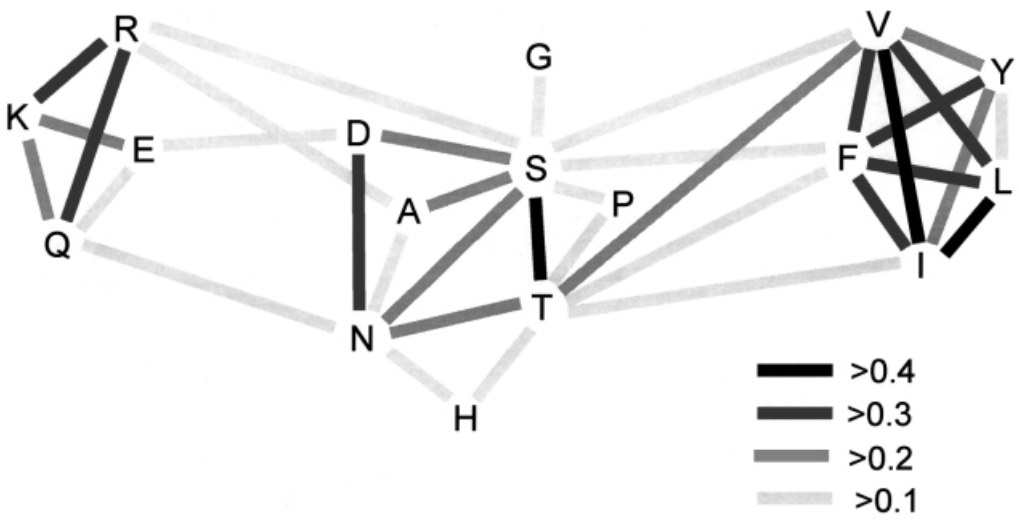

Fig. 3. Classification of amino acid residues according to similarity values of their surroundings. Darker lines join residues with higher degrees of similarity. Only residues with similarity score $m \geqslant 0.1$ are depicted.

Table I. Values of similarity of the amino acid residues surroundings (*m*)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	1.000	–0.168	0.029	–0.033	0.064	–0.023	0.014	0.068	0.031	0.140	–0.562	0.131	0.020	0.051	0.143	0.249	0.248	0.232	–0.234	0.011
C	–0.168	1.000	–0.406	–0.672	–0.082	–0.257	–0.220	–0.159	–0.527	–0.085	–0.787	–0.298	–0.298	–0.449	–0.392	–0.147	–0.156	–0.091	–0.464	–0.199
D	0.029	–0.406	1.000	0.109	–0.227	–0.034	–0.003	–0.336	0.036	–0.314	–0.839	0.329	0.005	0.0676	0.026	0.215	0.105	–0.182	–0.446	–0.203
E	–0.033	–0.672	0.109	1.000	–0.432	–0.299	–0.284	–0.465	0.224	–0.432	–0.877	–0.010	–0.218	0.183	0.090	–0.064	–0.145	–0.317	–0.461	–0.352
F	0.064	–0.082	–0.227	–0.432	1.000	–0.097	–0.016	0.380	–0.277	0.367	–0.529	–0.073	–0.067	–0.217	–0.113	0.143	0.148	0.345	0.021	0.352
G	–0.023	–0.257	–0.034	–0.299	–0.097	1.000	–0.118	–0.256	–0.239	–0.205	–0.806	0.044	0.044	–0.176	–0.119	0.120	–0.039	–0.080	–0.375	–0.178
H	0.014	–0.220	–0.003	–0.284	–0.016	–0.118	1.000	–0.130	–0.111	–0.087	–0.785	0.167	–0.031	0.0252	0.084	0.148	0.103	–0.055	–0.262	0.073
I	0.068	–0.159	–0.336	–0.465	0.380	–0.256	–0.130	1.000	–0.250	0.486	–0.452	–0.124	–0.187	–0.259	–0.139	–0.026	0.129	0.494	–0.021	0.237
K	0.031	–0.527	0.036	0.224	–0.277	–0.239	–0.111	–0.250	1.000	–0.236	–0.712	0.197	–0.171	0.287	0.387	0.028	–0.010	–0.167	–0.364	–0.203
L	0.140	–0.085	–0.314	–0.432	0.367	–0.205	–0.087	0.486	–0.236	1.000	–0.389	–0.123	–0.159	–0.180	–0.090	0.015	0.092	0.365	0.002	0.189
M	–0.562	–0.787	–0.839	–0.877	–0.529	–0.806	–0.785	–0.452	–0.712	–0.389	1.000	–0.682	–0.767	–0.742	–0.681	–0.610	–0.548	–0.514	–0.717	–0.644
N	0.132	–0.298	0.329	–0.010	–0.073	0.044	0.167	–0.124	0.197	–0.123	–0.682	1.000	0.063	0.161	0.183	0.271	0.241	–0.055	–0.309	–0.045
P	0.020	–0.298	0.005	–0.218	–0.067	0.044	–0.031	–0.187	–0.171	–0.159	–0.767	0.063	1.000	–0.083	–0.052	0.194	0.140	–0.034	–0.354	–0.119
Q	0.051	–0.449	0.068	0.183	–0.217	–0.176	0.025	–0.259	0.287	–0.180	–0.742	0.161	–0.083	1.000	0.375	0.085	–0.002	–0.143	–0.251	–0.112
R	0.143	–0.392	0.026	0.090	–0.113	–0.119	0.084	–0.139	0.387	–0.090	–0.681	0.183	–0.052	0.375	1.000	0.127	0.062	–0.037	–0.168	0.003
S	0.249	–0.147	0.215	–0.064	0.143	0.120	0.148	–0.026	0.028	0.015	–0.610	0.271	0.194	0.085	0.127	1.000	0.455	0.129	–0.249	0.052
T	0.248	–0.156	0.105	–0.145	0.148	–0.039	0.103	0.129	–0.010	0.092	–0.548	0.241	0.140	–0.002	0.062	0.455	1.000	0.247	–0.209	0.087
V	0.232	–0.091	–0.182	–0.317	0.345	–0.080	–0.055	0.494	–0.167	0.365	–0.514	–0.055	–0.034	–0.143	–0.037	0.129	0.247	1.000	–0.015	0.246
W	–0.234	–0.464	–0.446	–0.461	0.021	–0.375	–0.262	–0.021	–0.364	0.002	–0.717	–0.309	–0.354	–0.251	–0.168	–0.249	–0.209	–0.015	1.000	0.080
Y	0.011	–0.199	–0.203	–0.352	0.352	–0.178	0.073	0.237	–0.203	0.189	–0.644	–0.045	–0.119	–0.112	0.003	0.052	0.087	0.246	0.080	1.000

than to **Y** ($m = 0.352$), according to the data obtained. Since the only difference between the chemical structures of **F** and **Y** is the presence of the hydroxyl group, it becomes obvious that it is the influence of such a group that results in differences in surroundings of these residues in primary structures. It is noteworthy that aromatic amino acid residues considerably differ from each other by their surroundings and cannot be allocated into a separate group. Also, they cannot be totally included in the group of hydrophobic residues. Concise differentiation of hydrophobic residues from others is in good agreement with the suggestions about a leading role of hydrophobic interactions in the folding of a polypeptide chain (Pace, 1992; Rose and Wolfenden, 1993).

With the threshold value $m = 0.3$, the string **K–R–Q** (and **E** with threshold value $m = 0.2$) and the **D–N** pair emerge. The common property of both groups of residues is the ability to form hydrogen bonds. The major factor causing separation of these residues into two different groups is the size of the side chain. In this case low similarities between surroundings of **D** and **E** ($m = 0.109$), and **N** and **Q** ($m = 0.161$) are of particular interest. This could be accounted for by the major role of side chain size rather than similarity of side chain functionalities in the folding process.

Reduction of the threshold value m to 0.2 leads to the emergence of residues **A**, **V**, **D** and **E** in the nearest neighbourhood of residues **S** and **T**. All above-mentioned residues were assigned to the group of so-called 'residues with a small size of side chains' in earlier classifications. Cysteine residue (**C**) was included in the same group. However, on the basis of present data, **C** has a unique environment and cannot be included in any group.

The influence of the β-methyl group upon the value of similarity of amino acid residues surroundings is revealed by example of residues **S** and **T**. The presence of this group results in a higher degree of similarity in surroundings for a **T–V** pair ($m = 0.247$) as compared to **S–V** ($m = 0.129$). Thus, **T** takes an intermediate position between highly hydrophilic **S** and highly hydrophobic **V**. Similar differences are observed for pairs **S–A** and **T–A**, **S–I** and **T–I**, and others. It should be noted that the presence of beta methyl in **V**, **I** and **T** does not result in their allocation into a separate group.

With the threshold value $m = 0.1$, **S** and **T** have the greatest numbers of neighbours on the diagram (9 and 7, respectively). This multitude includes **P**, **G** and **H**, the surroundings of which have the least degrees of similarity with the surroundings of other residues. This fact suggests that **S** and **T** may substitute most residues in protein molecules with minimal effect upon local 3D structure.

Finally, there is a number of the amino acid residues (**M**, **C**, **P**, **G**, **H** and **W**), the surroundings of which have the least degree of similarity with the surroundings of other residues. Their uniqueness reflects the special role of these residues in formation of a protein structure. Thus, **M** is the leader residue almost in all native polypeptides. Residues **P** and **G** have allowed areas for torsion angles of the backbone, which differ essentially from those of other residues because of the unique organisation of proline side chain and the absence of side chain for glycine. Cysteine residues can form covalent bonds with distant segments of polypeptide chain. The tryptophane residue has the largest side chain, so its arrangement imposes specific requirements on the nearest neighbourhood. The side chain of **H** can participate in proton relay: this residue is frequently present in catalytic sites of enzymes. Substitution of any of these residues by any other is likely to result in disturbance of the local 3D structure of a protein.

With the threshold value $m < 0.1$, an overlap between groups of the residues is observed. Accordingly, consideration of lower levels of similarity of the surroundings is inexpedient.

In this study, a universal numerical measure of amino acid residues similarity based on the analysis of similarities of their surroundings in native protein sequences is elaborated. The classification of residues, based on this criterion, reveals essential differences from earlier classifications.

Similarity of chemical structure of side chains, such as aromaticity or presence of identical functional groups, has been demonstrated to be insufficient for allocation of the residues into groups, whereas the size of side chain can be foundational for such classification.

The concise differentiation of hydrophobic residues from others shows that hydrophobicity is the most important parameter of the amino acid residues, which influences the formation of 3D structure of protein.

Six amino acid residues having unique surroundings are revealed. The substitution of any of them by any other residue is likely to result in principle changes in local 3D organisation of a protein molecule with a high degree of probability.

The obtained results suggest that the criterion elaborated reflects structural features of amino acid residues. Thus, the proposed criterion as well as data about the environment of residues can be applied to evaluation of influence of amino acid substitutions on a 3D structure of proteins in studies utilising site-directed mutagenesis.

Acknowledgements

The authors are grateful to Ivan Yudushkin for his help in preparation of this manuscript.

References

Bordo,D. and Argos,P. (1991) *J. Mol. Biol.*, **217**, 721–729.
Cserzo,M. and Simon,I. (1989) *Int. J. Pept. Protein Res.*, **34**, 184–195.
Hurley,J.H., Baase,W.A. and Matthews,B.H. (1992) *J. Mol. Biol.*, **224**, 1143–1159.
Janin,J. (1979) *Nature*, **277**, 491–492.
Johnson,M.S. and Overington,J.P. (1993) *J. Mol. Biol.*, **233**, 716–738.
Kyte,J. and Doolite,R. (1982) *J. Mol. Biol.*, **157**, 105–132.
Lim,W.A., Farruggio,D.C. and Sauer,R.T. (1992) *Biochemistry*, **31**, 4324–4333.
Murphy,L.R., Wallqvist,A. and Levy,R.M. (2000) *Protein Eng.*, **13**, 149–152.
Pace,C.N. (1992) *J. Mol. Biol.*, **226**, 29–35.
Poroykov,V.V., Esipova,N.G. and Tumanyan,V.G. (1976) *Mol. Biophys. (Moscow)*, **21**, 397–400 (in Russian).
Rose,G., Geselowitz,A., Lesser,G., Lee,R. and Zehfus,M. (1985) *Science*, **229**, 834–838.
Rose,G.D. and Wolfenden,R. (1993) *Annu. Rev. Biophys. Biomol. Struct.*, **22**, 381–409.
Taylor,W.R. (1986) *J. Mol. Biol.*, **188**, 233–258.
Topham,C.M., Srinivasan,N. and Blundell,T.L. (1997) *Protein Eng.*, **10**, 7–21.
Wolfenden,R., Andersson,L., Cullis,P. and Southgate,C. (1981) *Biochemistry*, **20**, 849–855.
Zhang,X.-J., Baase,W.A. and Matthews,B.W. (1992) *Protein Sci.*, **1**, 761–776.

Received August 21, 2000; revised February 23, 2001; accepted March 12, 2001